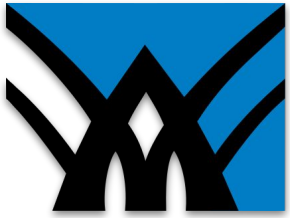
Sächsische Akademie der Wissenschaften zu Leipzig

UNIVERSITÄT LEIPZIG

InfAI Institut für Angewandte Informatik



WORTSCHATZ LEIPZIG

Working with (English News) Corpora at the Leipzig Corpora Collection

Thomas Eckart (eckart@saw-leipzig.de)
Felix Helfer (helfer@saw-leipzig.de)
Erik Körner (koerner@saw-leipzig.de)

Project Wortschatz Leipzig

Wortschatz Leipzig

Short project history

- Started in 1994, first Web portal in 1998, first API since 2004
- Text databases with more than 30 billion sentences in more than 250 languages
- Strong focus on “Web as a corpus” paradigm
- Cooperation between Leipzig University, SAW and InfAI

Competences

- Handling large amounts of textual content
- Investing in algorithms instead of individual data sets
- Development of procedures independent of individual languages to internationalize with minimal effort

NFDI – Text+

In Zusammenarbeit mit nfdi Nationale Forschungsdaten Infrastruktur

Text+

Participating in the Text+ project via the *Saxon Academy of Sciences and Humanities*

Building on long experience in providing linguistic resources to users in

- Data domain “lexical resources”
  - Providing our own dictionaries in the NFDI
  - Working on interfaces and search engines for accessing lexical resources
- Task area “infrastructure / operations”
  - Findability & Interoperability of text-based resources in a federated infrastructure
  - *Federated Content Search* (FCS) as central search component in Text+
  - (Linguistic) Linked Open Data

Wortschatz Leipzig – Thematic focus

Text analysis

- Compiling large corpora as text databases
- Use of a complex preprocessing pipeline
- Focus on statistical analysis:
  - Frequency- and cooccurrences-based analysis, sentiment analysis, topic modelling, document classification

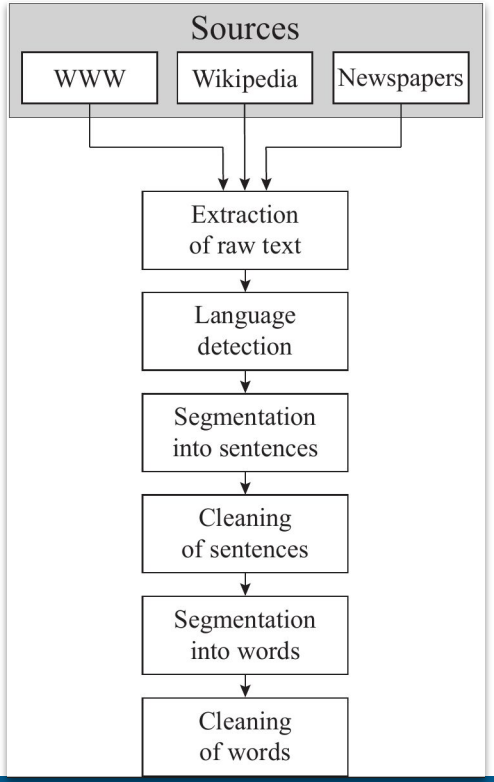
Applications

- Support with dictionary work and ontology creation
- Knowledge management / information retrieval: extraction of connections & keywords etc.

Wortschatz Leipzig – Sources & Preprocessing

- Online news sites
  - RSS-Feeds (daily, in ca. 80 languages)
- Web text
  - Random crawling (of selected TLDs)
  - Stopword-driven (BootCat)
  - User-driven crawling
- Wikipedia
- Administrative texts, …

- Pipeline for crawling, cleaning, language-identification, statistical analysis, publication



Web data

Crawling since 2002; over a trillion crawled HTML-pages so far.

250+ languages available online - for an overview, see the language selector on:

https://corpora.wortschatz-leipzig.de/en

Example: English news corpus 2023*

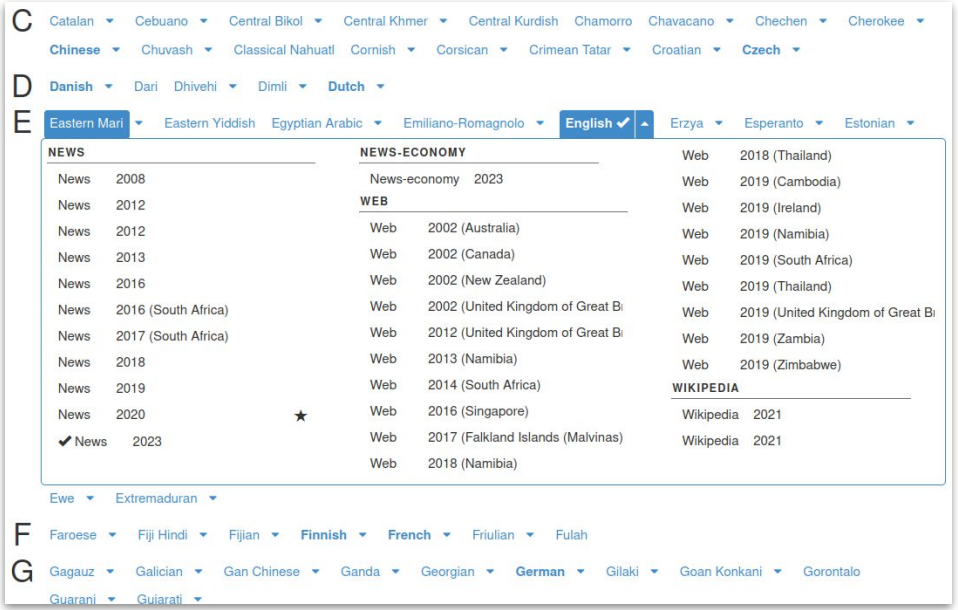
- Based on daily RSS feeds
- Around 3,7M documents from 3400 news sites in English language all over the globe (66% .com, 17% .uk, 3% .au, 2,8% .net,...)
- Around 70M sentences remaining after standard toolchain processing and cleaning
- Annotations:
  - POS tagged
  - POS tag based sentence signatures
  - Word cooccurrences (windows: sentence & ± 1 token)
  - Cooccurrences-based word embeddings
  - …

* up to 2023-11-19

Applications

Corpus portal

- Idea: distant reading approach to corpora, condensed presentation of linguistic and statistical analysis
- References to other WS portals
- Only selection of available data



→ https://corpora.wortschatz-leipzig.de/en

Corpus portal – German News 2022 – *Universität*

Frequency

Linguistic information

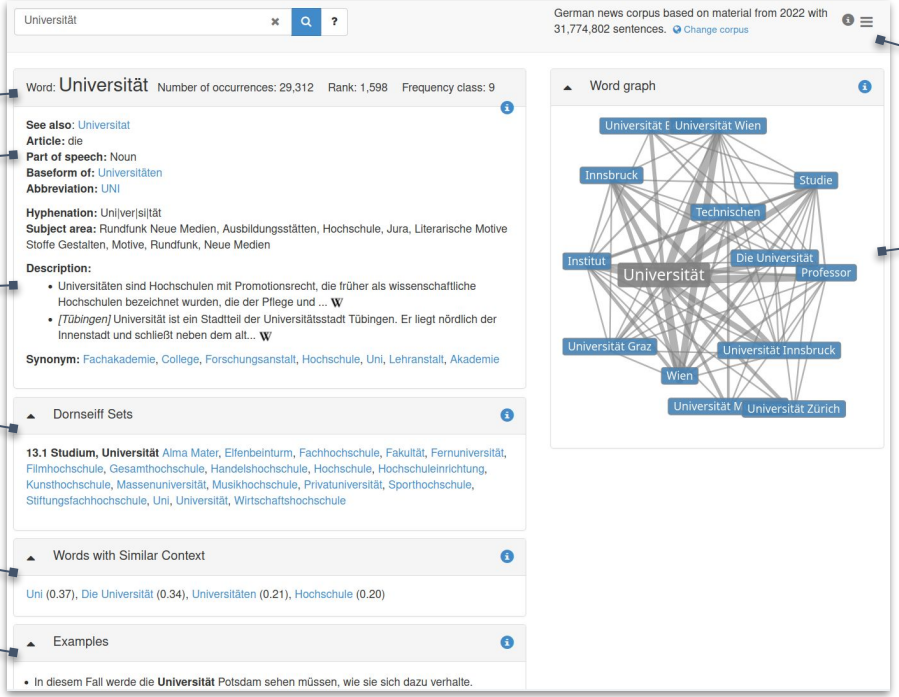
Definitions

Dornseiff

Embedding similarity

Examples

[...]



Corpus metadata & Links

Cooccurrences graph

→ https://corpora.wortschatz-leipzig.de/en?corpusId=deu_news_2022&word=Universität

Corpus portal – English News 2023 – *values*

Frequency

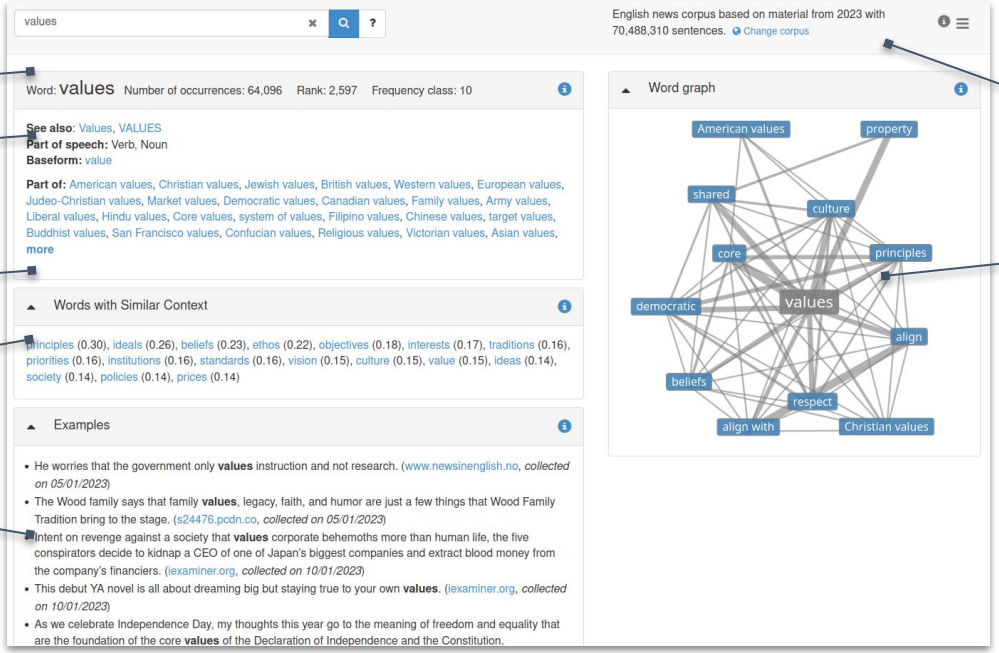
Various linguistic information

String similarity

Embedding similarity

Examples

[Word co-occurrences]



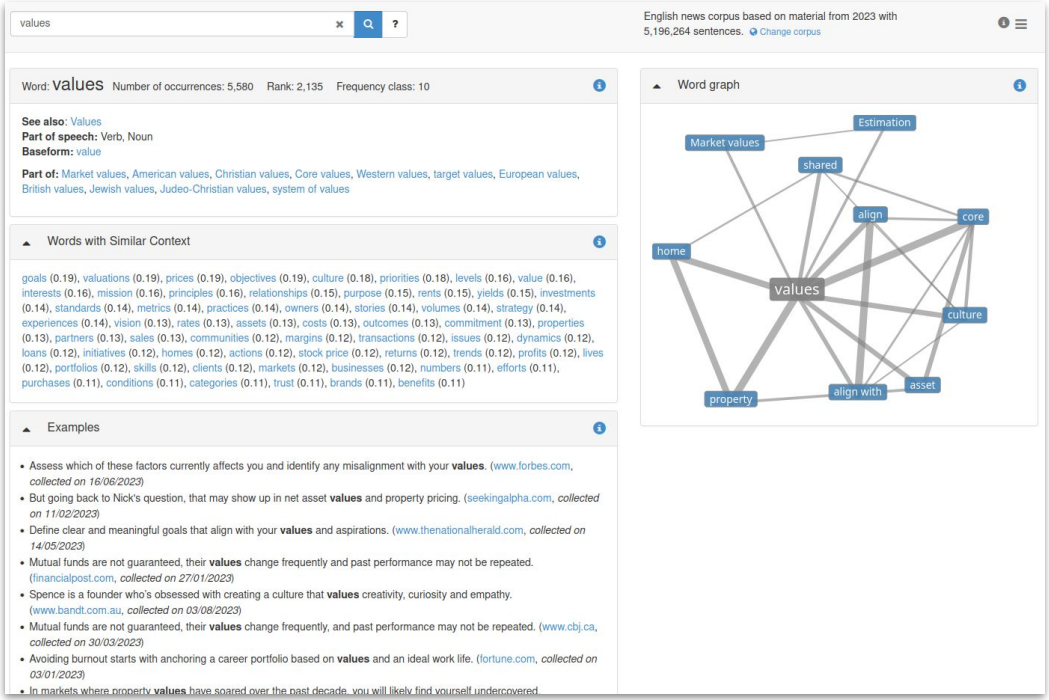
Corpus metadata & Links

Cooccurrences graph

→ https://corpora.wortschatz-leipzig.de/en?corpusId=eng_news_2023&word=values

Corpus portal – Subcorpus “economy” – *values*

- Filtered subcorpus, part of *English news 2023*
- Document filter based on simple keyword extraction
- Size:
  - 197K documents
  - 5.2M sentences
  - 111M word tokens



→ https://corpora.wortschatz-leipzig.de/en?corpusId=eng_news-economy_2023&word=values

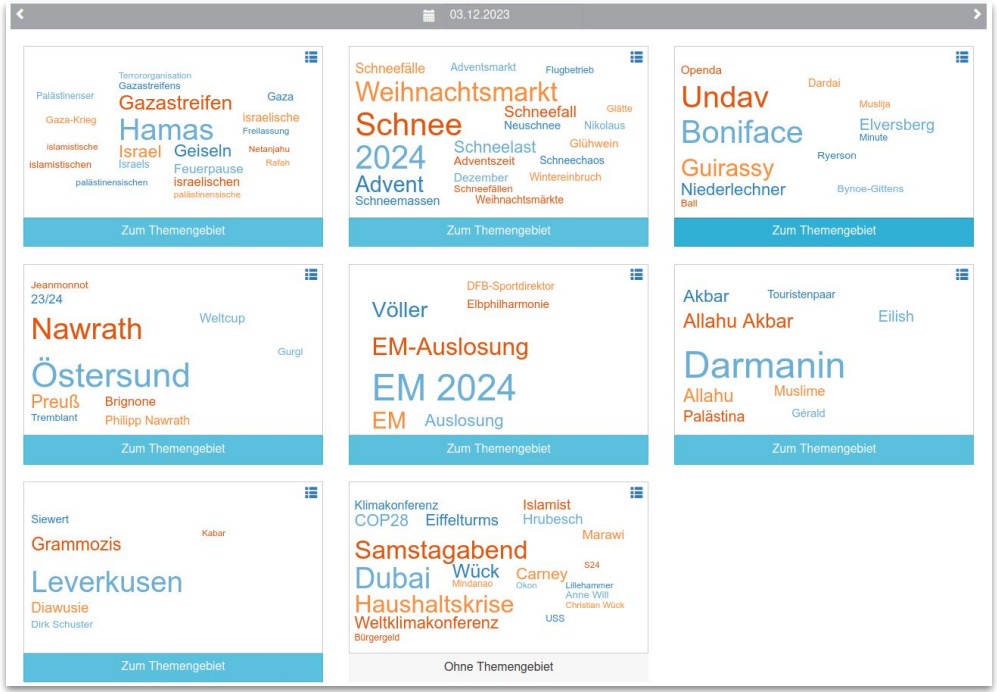
“Words of the day”

WORDS of the DAY
LEIPZIG UNIVERSITY

- Idea: daily news text can be used to compare thematic developments and identify salient topics on a daily basis
- Topic Modelling (based on LDA using a standard library)
- Material:
  - German: ca. 8,000 articles per day (75% .de, 15% .at, 4% .ch)
  - English: ca. 16,000 articles per day (worldwide)
- Presentation at “words of the day” portal (currently only in German language)

→ https://wod.wortschatz-leipzig.de/

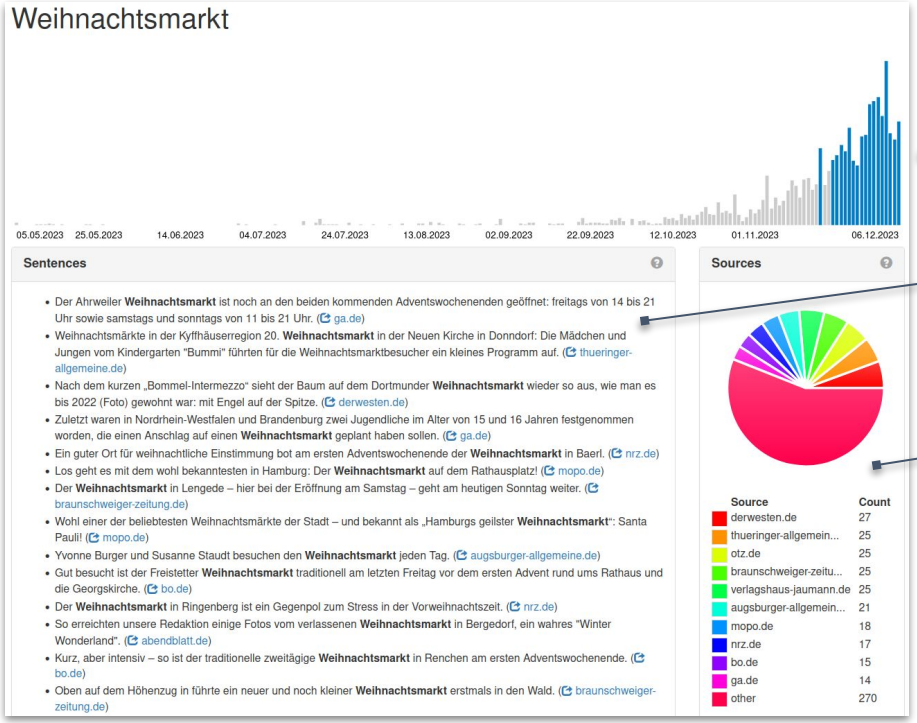
Topics of the day



2023-12-03

→ https://wod.wortschatz-leipzig.de/

Diachronic analysis of words – *Christmas market*



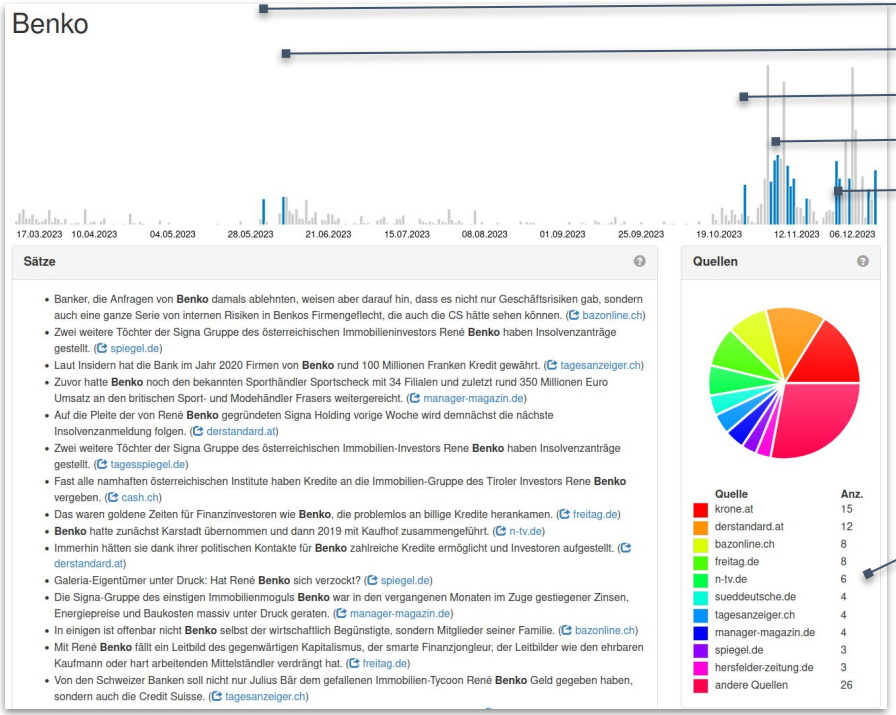
Relative frequency over time

References

Sources

→ Weihnachtsmarkt

Diachronic analysis of words – *Benko*



Signa Holding sells *kika/Leiner*

Insolvency of *kika/Leiner*

Insolvency of *Signa Sports United*

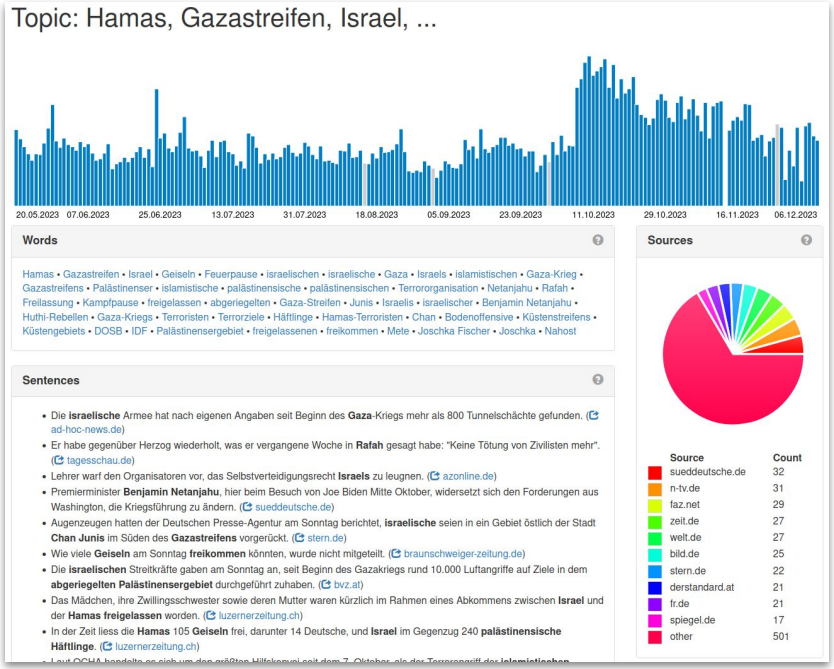
Leaves *Signa Holding*

Insolvency of *Signa Holding*

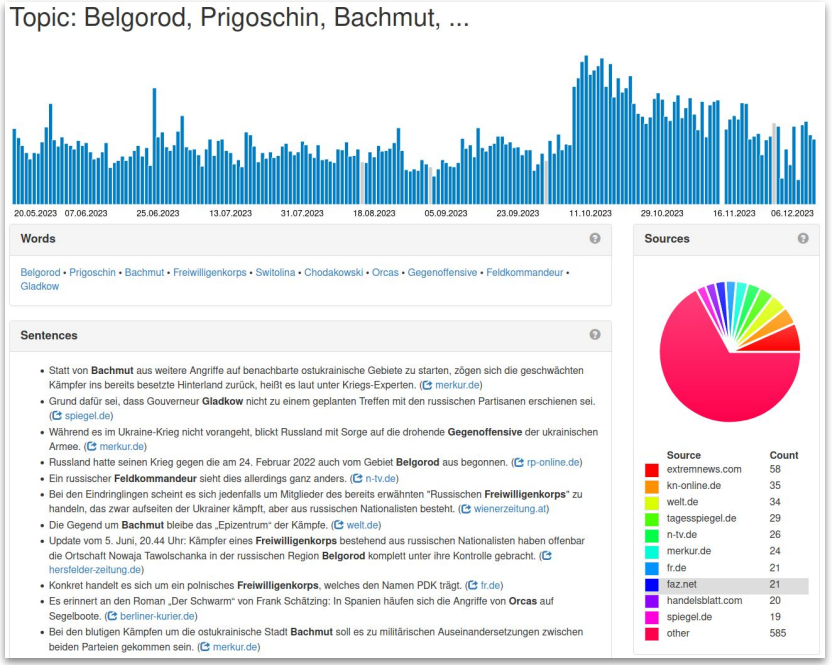
Quelle	Anz.
krone.at	15
derstandard.at	12
bazonline.ch	8
freitag.de	8
n-tv.de	6
sueddeutsche.de	4
tagesanzeiger.ch	4
manager-magazin.de	4
spiegel.de	3
hersfelder-zeitung.de	3
andere Quellen	26

→ Benko

Diachronic development of topics – topic "*war*"



→ 2023-12-03



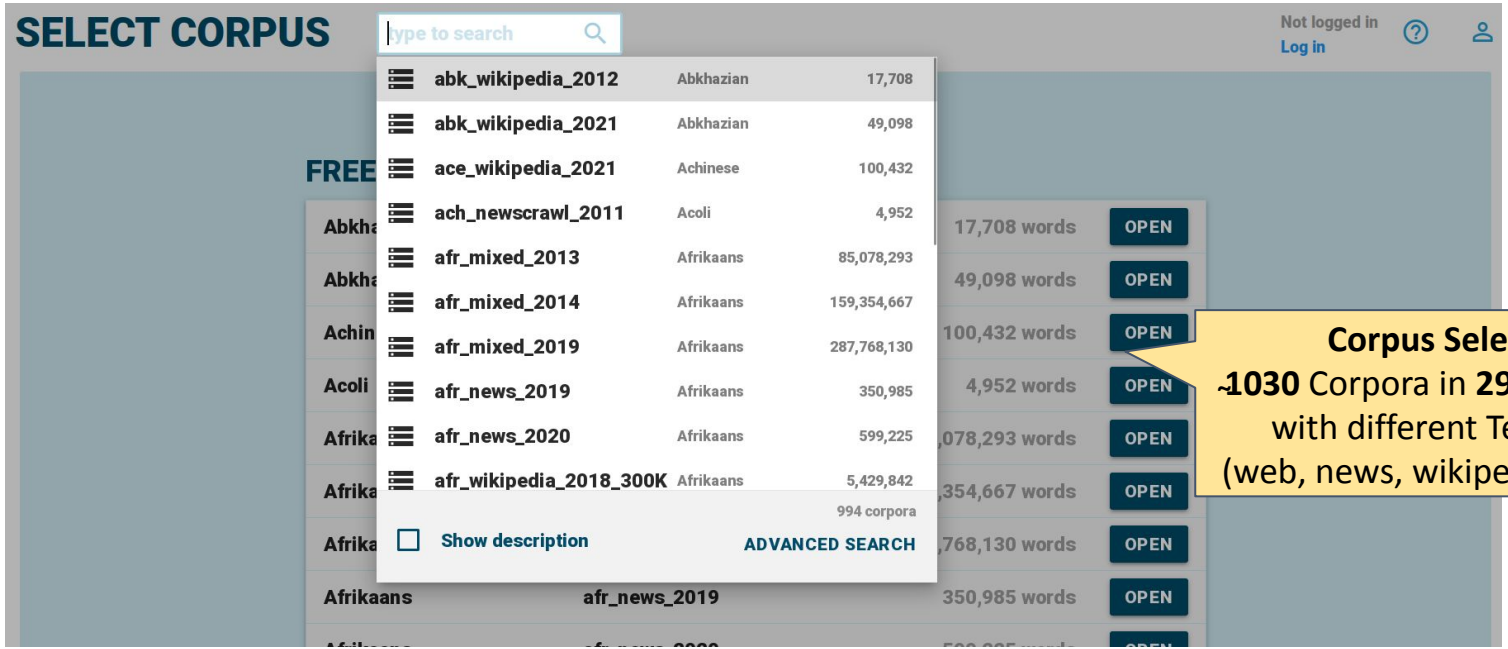
→ 2023-06-05

Corpus query engine NSE

- Idea: computer linguistics has long-term experience in working with large text collections
- Various tools to support work (filtering, selection, terminology work, sub-corpora comparisons etc.)
- One prominent application: NoSketchEngine (NSE)
  - Open-source version of a commercial tool, that focuses on lexicographers (“word sketches”)
  - Very efficient indexing of **large** text corpora
  - Allows: Filterung, dynamic sub-corpora generation, keyword extraction, corpus comparison, flexible word collocation presentation → distant **and** close reading approaches

→ https://cql.wortschatz-leipzig.de/

Corpus query engine NSE



Corpus Selection
~**1030** Corpora in **293** Languages
with different Text Types
(web, news, wikipedia, private)

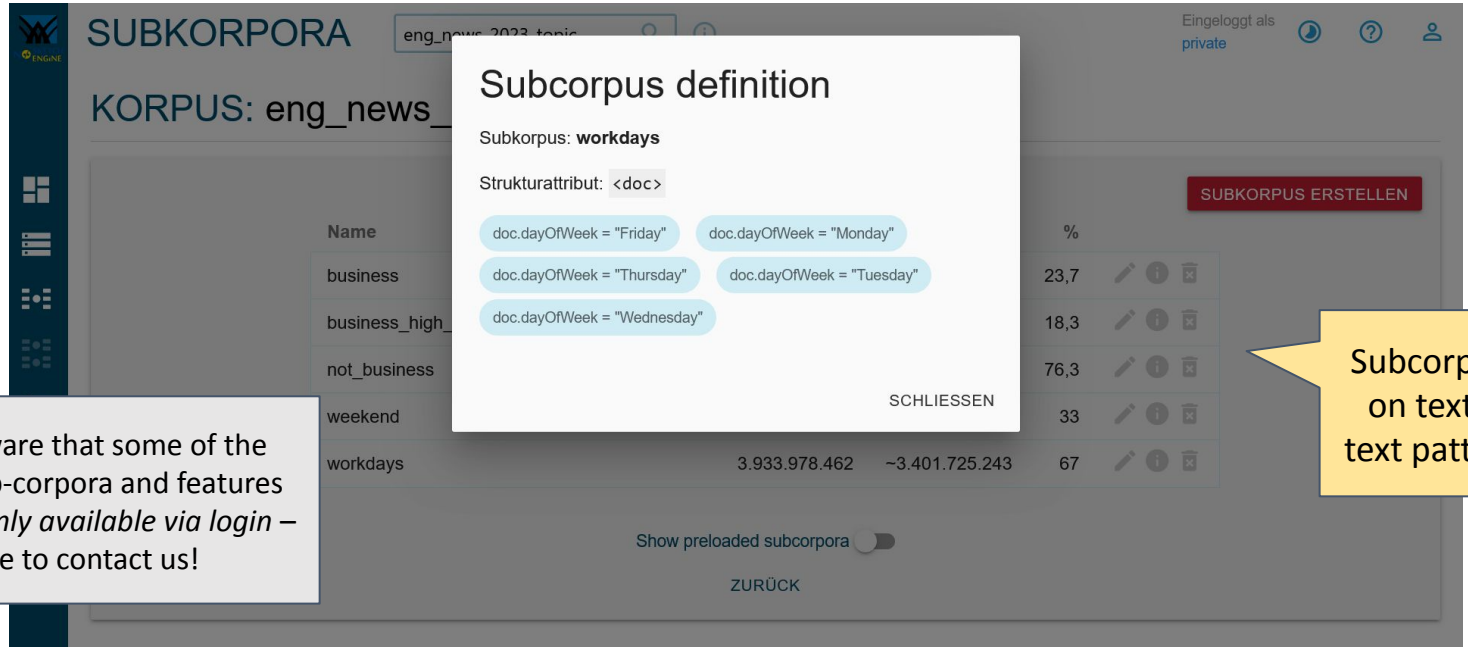
→ https://cql.wortschatz-leipzig.de/#corpus?tab=advanced

Corpus query engine NSE

- Example with extensive linguistic annotation: “eng_news_2023_topic”
- Processing of RSS/Feed based daily news crawl dataset with
  - **Document: LCC** html-to-text extraction and document language filtering
  - **Document: LCC** Sentence Segmentation and Tokenization
  - **Sentence: TextBlob** 🔗 Sentiment Polarity and Subjectivity prediction
  - **Sentence: Transformers** 🔗 AGNews topic classification
  - **Token: spaCy** 🔗 for linguistic annotation, e.g., lemmatization, *part-of-speech tagging, morphological features*, word shape, named entity recognition
- Statistics:
  - 5 Mio. documents, 417 Mio. sentences, 5.9 Bn. tokens

→ https://cql.wortschatz-leipzig.de/#dashboard?corpname=eng_news_2023_topic&corp_info=1
https://cql.wortschatz-leipzig.de/#dashboard?corpname=eng_news_2023_topic2&corp_info=1

Subcorpus definition and creation

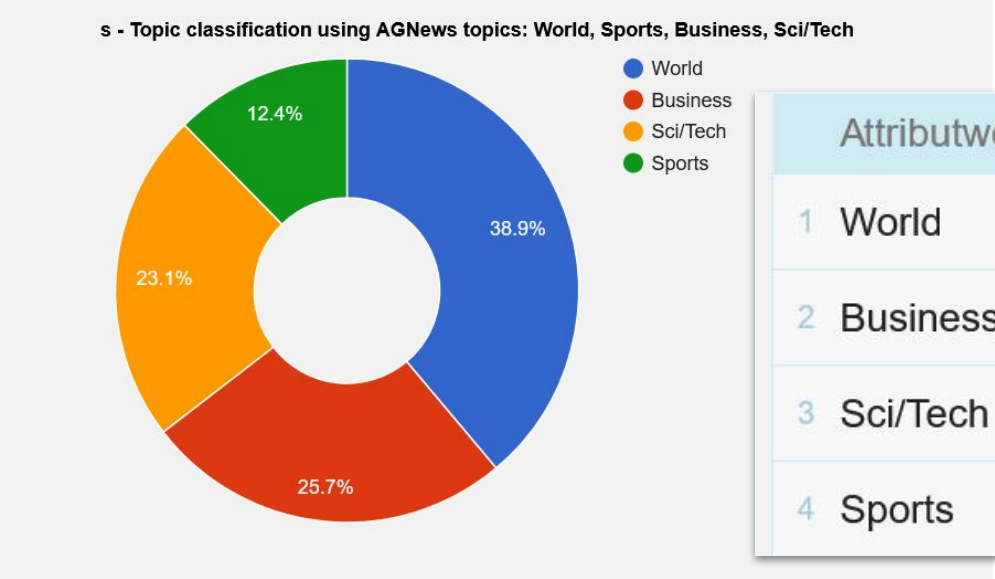


Please be aware that some of the presented sub-corpora and features are *currently only available via login* – Feel free to contact us!

Subcorpora based on text types or text patterns (CQL)

→ https://cql.wortschatz-leipzig.de/#ca-subcorpora?corpname=eng_news_2023_topic

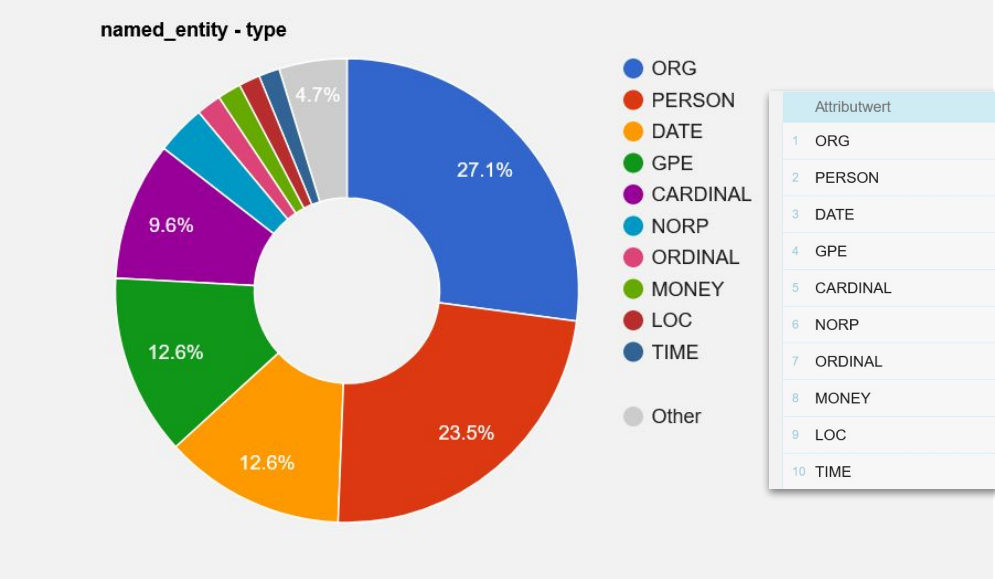
Distribution of sentences by AGNews topic



	Attributwert	Structure frequency ?	
1	World	161.891.852	•••
2	Business	107.013.019	•••
3	Sci/Tech	96.120.559	•••
4	Sports	51.545.339	•••

→ https://cql.wortschatz-leipzig.de/#text-type-analysis?corpname=eng_news_2023_topic&tab=basic&filter=containing&wlattr=s.topic …

Distribution of Named Entities

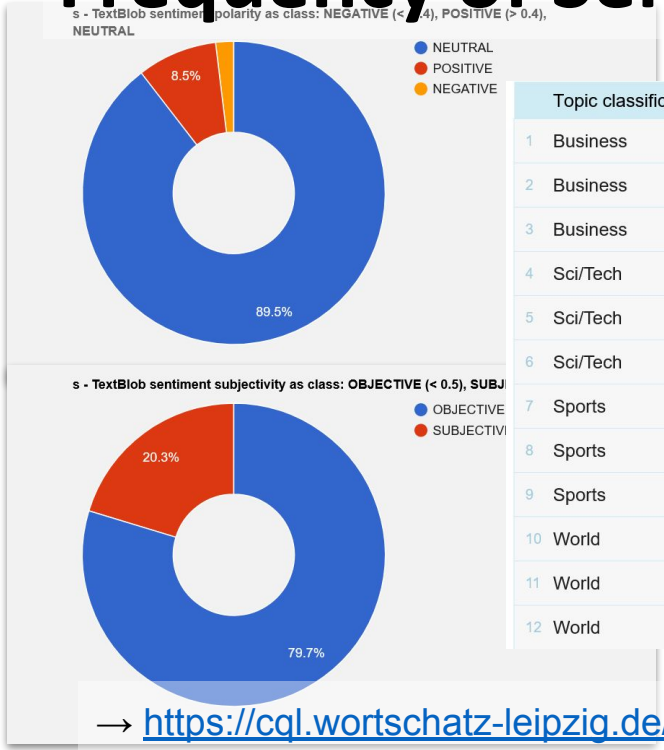


	Attributwert	Structure frequency
1	ORG	161.526.665
2	PERSON	140.127.419
3	DATE	75.246.554
4	GPE	75.192.146
5	CARDINAL	57.477.561
6	NORP	20.556.376
7	ORDINAL	10.293.511
8	MONEY	9.805.882
9	LOC	8.847.620
10	TIME	8.700.431

	Attributwert	Structure frequency
11	FAC	7.295.712
12	PRODUCT	6.622.137
13	WORK_OF_ART	6.195.678
14	EVENT	3.920.734
15	QUANTITY	1.890.060
16	LAW	1.226.776
17	PERCENT	636.034
18	LANGUAGE	452.398

→ https://cql.wortschatz-leipzig.de/#text-type-analysis?corpname=eng_news_2023_topic&tab=basic&filter=containing&onecolumn=1&wlattr=named_entity.type

Frequency of Sentiment Polarity by Topic

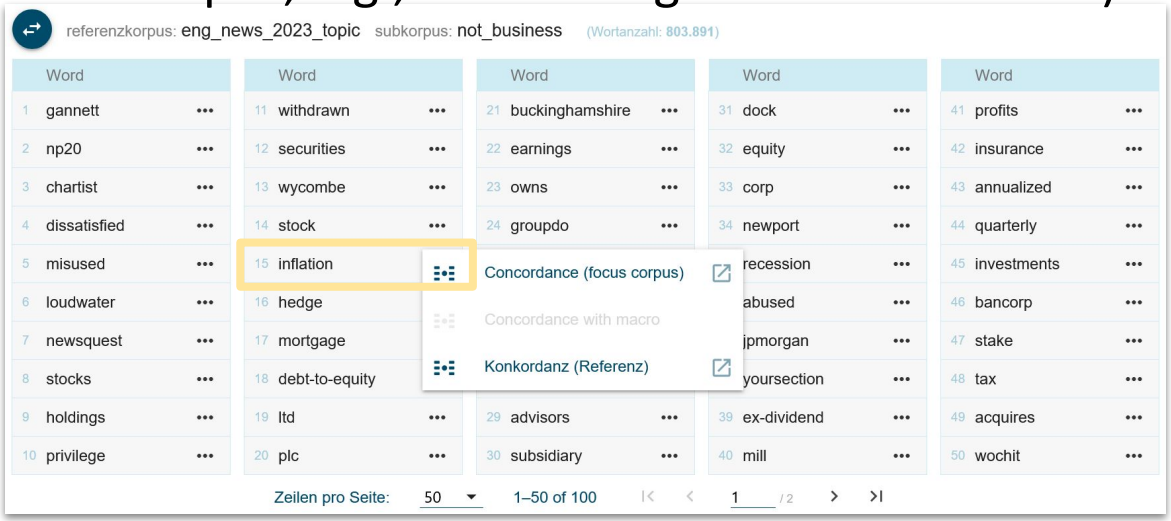


	Topic classification using AGNews topics	TextBlob sentiment polarity	Frequenz
1	Business	NEGATIVE	4.892
2	Business	NEUTRAL	1.339.940
3	Business	POSITIVE	94.274
4	Sci/Tech	NEGATIVE	19.387
5	Sci/Tech	NEUTRAL	1.604.824
6	Sci/Tech	POSITIVE	74.020
7	Sports	NEGATIVE	4.916
8	Sports	NEUTRAL	344.074
9	Sports	POSITIVE	14.612
10	World	NEGATIVE	12.307
11	World	NEUTRAL	1.459.671
12	World	POSITIVE	82.111

→ https://cql.wortschatz-leipzig.de/#concordance?corpname=eng_news_2023_topic&queryselector=cql&attrs=word&viewmode=kwic&attr_allpos=all&refs_up=0& ...

Automatic keyword extraction

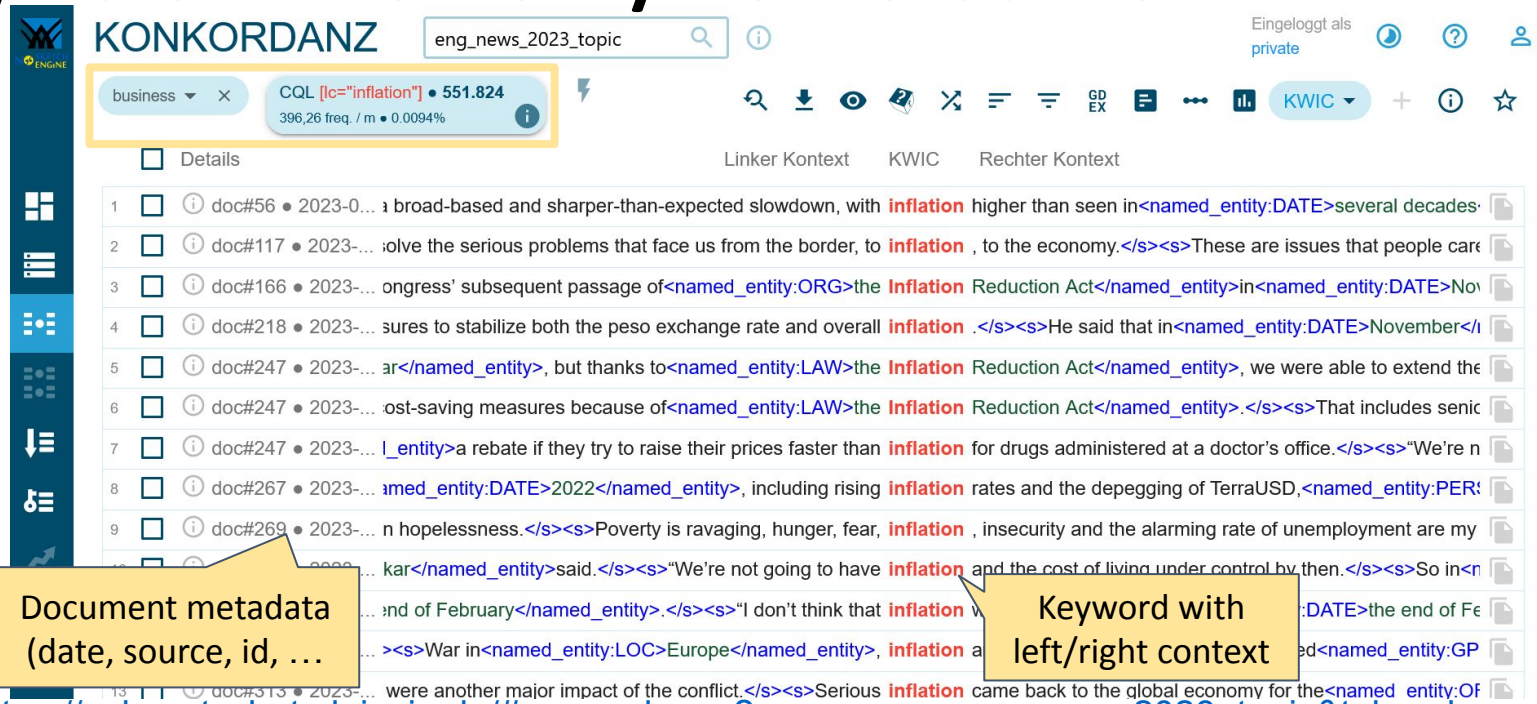
- Comparing word frequencies using a reference corpus (or subcorpus, e.g., business against non-business)



	Word (lowercase)	Frequency per million? Focus	Reference
1	municipality	107.30	0.76
2	railroad	63.60	0.66
3	ncaa	43.40	0.19
4	genus	90.89	1.48
5	ohio	71.07	1.25
6	pennsylvania	80.10	1.56
7	illinois	64.56	1.23
8	michigan	72.81	1.59
9	missouri	43.40	0.61
10	baseball	87.21	2.34
11	kansas	44.06	0.72
12	billboard	48.95	0.91
13	voivodeship	25.60	0.02
14	carolina	78.31	2.06
15	gmina	24.80	0.00
16	wisconsin	42.29	0.70
17	nfl	45.96	0.85
18	touchdown	31.37	0.29
19	texas	118.54	3.91
20	tennessee	41.85	0.84
21	minnesota	44.23	0.96
22	province	138.31	5.06
23	ontario	55.84	1.48
24	oklahoma	33.91	0.54
25	criticized	34.58	0.57

→ https://cql.wortschatz-leipzig.de/#keywords?corpname=eng_news_2023_topic&tab=advanced&k_attr=word&k_itemsPerPage=50&ktab=keywords& …

Keyword-in-Context / Concordance



→ https://cql.wortschatz-leipzig.de/#concordance?corpname=eng_news_2023_topic&tab=advanced&queryselector=cql&cql=%5Blc%3D%22inflation%22%5D&usesubcorp=business&showresults=1

Statistics / Correlations

- Correlations to word “***inflation***” over time, by topic or source

	Domain	Frequenz
1	seekingalpha.com	29.832
2	forextv.com	22.875
3	dailymail.co.uk	21.377
4	forexlive.com	17.775
5	marketscreener.com	8.637
6	businessmirror.com.ph	8.574
7	business.inquirer.net	8.110
8	biztoc.com	7.619

	Doc.dayOfWeek	Frequenz
1	Wednesday	111.557
2	Thursday	103.073
3	Tuesday	88.723
4	Friday	77.314
5	Sunday	61.097
6	Monday	60.128
7	Saturday	49.932

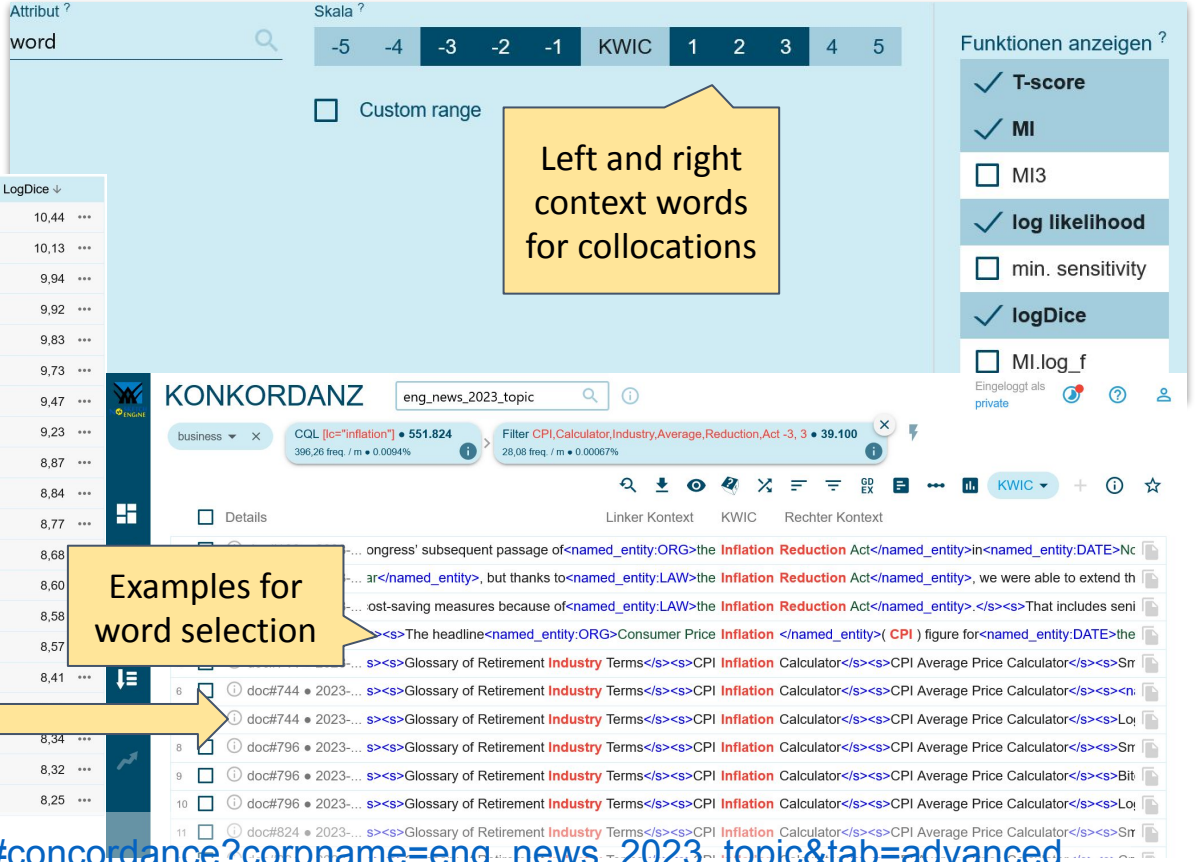
	Doc.month	Frequenz ↓
1	January	80.121
2	February	67.961
3	March	63.825
4	June	57.068
5	May	54.878
6	April	52.990
7	August	45.315
8	September	
9	July	
10	October	
11	November	20.124

Work on subset with filter on certain month

	Topic classification using AGNews topics	Frequenz
1	Business	551.824

→ https://cql.wortschatz-leipzig.de/#concordance?corpname=eng_news_2023_topic&tab=advanced&queryselector=cql&attrs=word&viewmode=kwic&attr_allpos=all&refs_up=0&shorten_refs=1& …

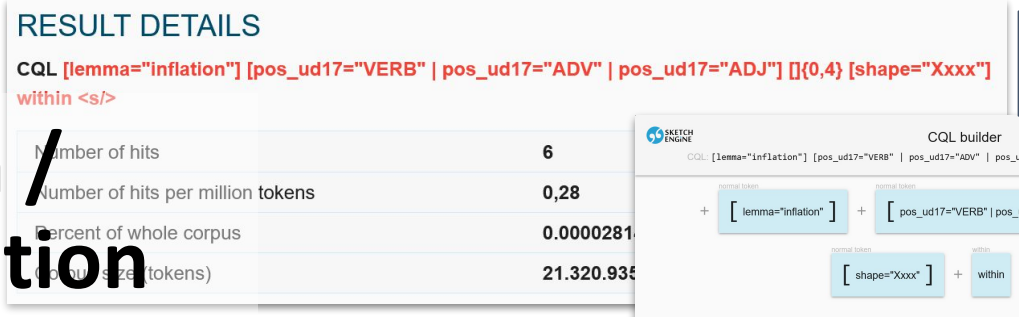
Collocations

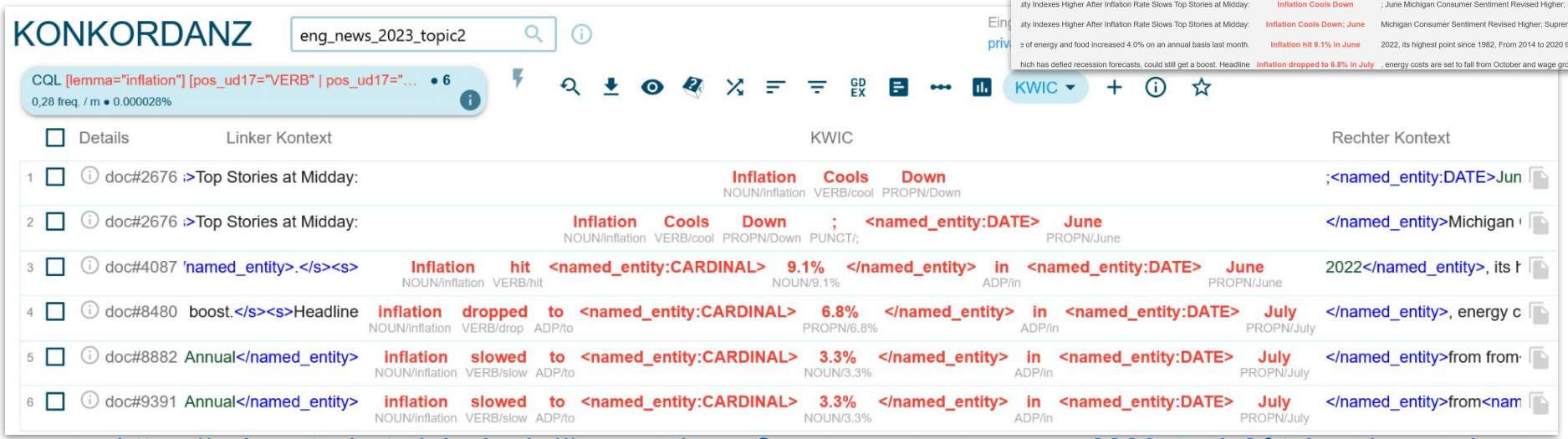


		Word	Kookkurrenzen ?	Kandidaten ?	T-score	MI	Log likelihood	LogDice ↓	
1	✓	CPI	26.322	67.167	162,20	12,03	399.501,51	10,44	•••
2	✓	Calculator	20.990	60.701	144,84	11,85	311.814,80	10,13	•••
3	✓	Industry	20.942	144.996	144,62	10,59	269.470,98	9,94	•••
4	✓	Average	21.037	161.451	144,94	10,44	266.039,94	9,92	•••
5	☐	rate	28.260	467.262	167,85	9,33	312.320,25	9,83	•••
6	☐	high	28.922	561.491	169,75	9,10	310.130,69	9,73	•••
7	☐	Reduction	12.311	18.721	110,94	12,77	204.536,34	9,47	•••
8	☐	Act	12.100	109.830	109,91	10,20	148.497,45	9,23	•••
9	☐	rates	13.913	423.109	117,62	8,45	136.074,63	8,87	•••
10	☐	food	14.654	495.677	120,67	8,30	140.182,77	8,84	•••
11	☐	rising	9.287	143.245	96,23	9,43	103.642,30	8,77	•••
12	☐	down	15.736	705.025	124,92	7,89	141.634,28	8,68	•••
13	☐	data	9.750	271.288	98,48	8,58	97.043,89	8,60	•••
14	☐	Terms	20.940	1.237.272	143,90	7,49	177.024,53	8,58	•••
15	☐	halve	6.487	8.796	80,53	12,94	110.257,50	8,57	•••
16	☐	fight	6.536	79.780	80,75	9,77	76.080,88	8,41	•••
17	☐	core							
18	☐	higher	8.883	346.155	93,90	8,09	82.341,30	8,34	•••
19	☐	interest	9.417	415.656	96,64	7,91	84.934,00	8,32	•••
20	☐	bring	6.922	193.174	82,98	8,58	68.818,73	8,25	•••

→ https://cql.wortschatz-leipzig.de/#concordance?corpname=eng_news_2023_topic&tab=advanced&queryselector=cql&attrs=word&viewmode=kwic&attr_allpos=all&refs_up=0&shorten_refs=1& …

Pattern search / Token Information

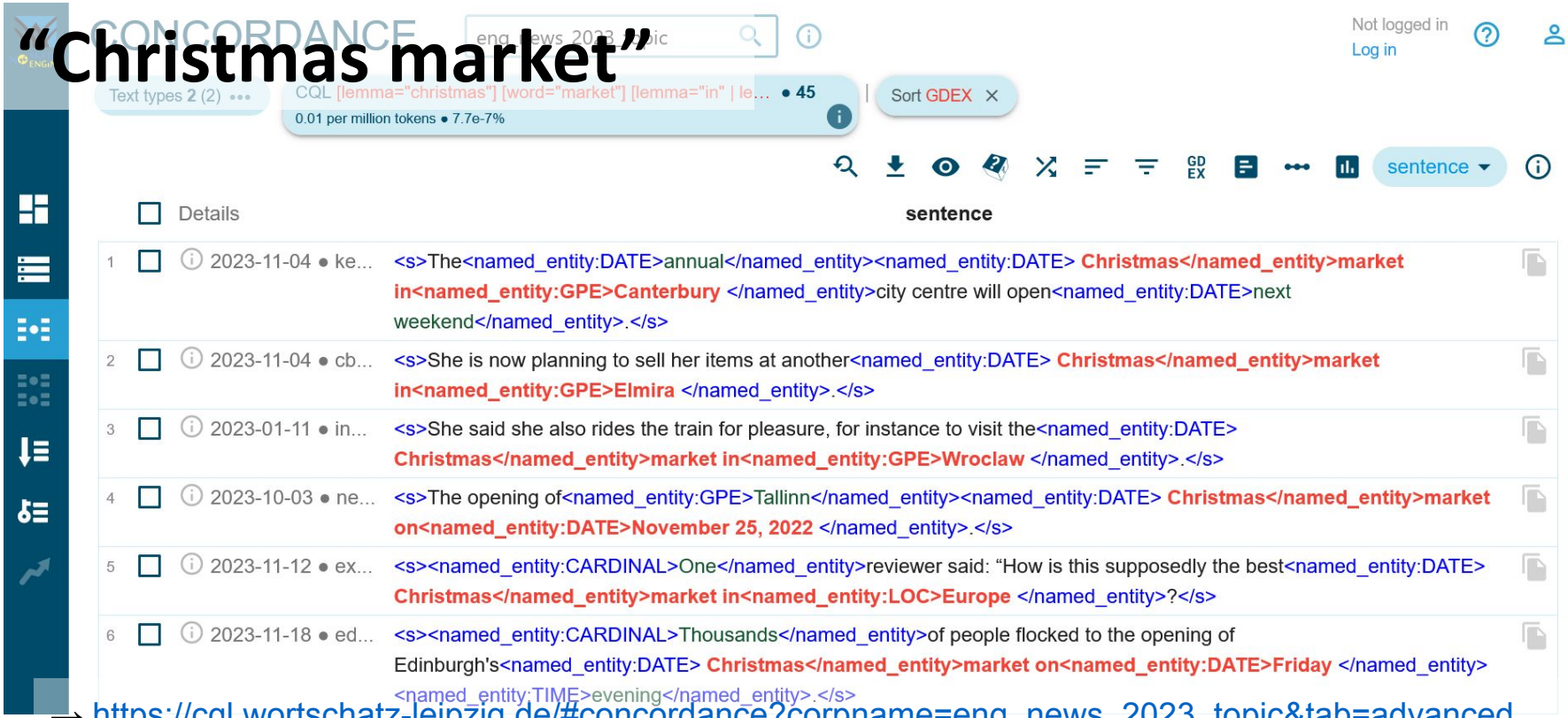




→ https://cql.wortschatz-leipzig.de/#concordance?corpname=eng_news_2023_topic2&tab=advanced&queryselector=cql&attrs=word&viewmode=kwic&attr_allpos=all&refs_up=0&shorten_refs=1& …

→ https://cql.wortschatz-leipzig.de/#concordance?corpname=eng_news_2023_topic2&tab=advanced&queryselector=cql&attrs=word%2Cpos_ud17%2Clemma%2Cpos&viewmode=sen&attr_allpos=kw …

“Christmas market”



→ https://cql.wortschatz-leipzig.de/#concordance?corpname=eng_news_2023_topic&tab=advanced&queryselector=cql&keyword=war&attrs=word&viewmode=sen&attr_allpos=all&refs_up=0&...

Use-Cases and Cooperations

Wortschatz Leipzig – User Groups

Typical **user groups** include:

- Philologists, linguists, computer linguists
- Researchers in social science and other text-oriented domains
- NLP & digital humanities
- Translators, authors, journalists, …
- General public interested in text generation
- …

WestAI

- One of four German “AI service centers” (*Center West*), funded by the **BMBF**.

- Important focus: the creation and provision of **large, multimodal AI models**.

  → Delivery of a **large collection of high-quality Wortschatz corpora** from 2021-2022 to WestAI and its partnering institutes and projects.

  → The data are to be used for text mining and the development of **German large language models**.

https://westai.de

LLM Training at ScaDS.AI

- *Center for Scalable Data Analytics and Artificial Intelligence Dresden/Leipzig*

- Collaboration for a research project to train a state-of-the-art **Large Language Model (LLM) for German**

- The *Wortschatz Leipzig project* can deliver the **large amount of raw text** necessary for training.

- In addition, personnel can lend their **expertise** in **web crawling** and **text/linguistic preprocessing**

Project WINTER

- **EU-funded project**, which aims to develop a web interactive platform for the management of coal regions in Germany, Greece and Poland.

- **26 Wortschatz news corpora for German, Greek and Polish** were used by the TH Georg Agricola (THGA) for an analysis of **media representation and perception** of the structural reduction of coal industry in the different countries.

- Utilized the **Wortschatz portal** as an intuitive, easily usable tool.

https://winter-project.eu

Sentiment analysis on customer reviews

- Extensive stock of product reviews as a corpus for identifying customers' sentiment about various product characteristics
- *SentiWS* containing 34K German words with their sentiment polarity ([-1,+1]) (German language only) combined with word embeddings for a larger recall
- Identifying
  - Positive/negative characteristics of single products
  - Summary of negative characteristics of product groups
  - Positive/negative characteristics of products in comparison with their product groups

→ https://wortschatz-leipzig.de/en/download#sentiWSDownload

SFB 1199 “Processes of Spatialization under the Global Condition”

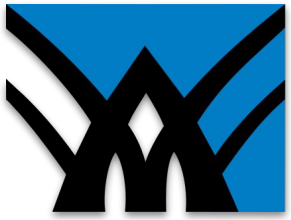
- Processing of historical text data
  - Texts of geographical societies of three countries from the 19th and 20th centuries
  - Collections of magazines and books from the 20th century in different languages
  - Pre-revolutionary French dictionaries
- Corpus analysis and terminology extraction
- Approach: distant reading combined with close reading “drill-down”
- Analysis of vocabulary change

Other users

… include **private corporations**,

… **researchers**, from a multitude of institutions

… and **you**?



WORTSCHATZ LEIPZIG

https://wortschatz-leipzig.de